Available at WWW.MATHEMATICSWEB.ORG
POWERED BY SCIENCE DIRECT
Journal of Multivariate Analysis 86 (2003) 213–226
Journal of Multivariate Analysis
http://www.elsevier.com/locate/jmva

Efficient estimation in conditional single-index regression

Michel Delecroix,[a] Wolfgang Härdle,[b] and Marian Hristache[a,*]

[a] ENSAI and CREST, Rue Blaise Pascal, Campus de Ker Lann 35170 Bruz, France
[b] Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin, Spandauer Str. 1, D-10178 Berlin, Germany

Received 12 May 1998

Abstract

Semiparametric single-index regression involves an unknown finite-dimensional parameter and an unknown (link) function. We consider estimation of the parameter via the pseudo-maximum likelihood method. For this purpose we estimate the conditional density of the response given a candidate index and maximize the obtained likelihood. We show that this technique of adaptation yields an asymptotically efficient estimator: it has minimal variance among all estimators.

AMS 2000 subject classifications: primary 62G05; secondary 62G20, 62G07, 62G08

Keywords: Single-index model; Pseudo-maximum likelihood; Semiparametric efficiency bound

1. Introduction

A single-index response model has the form

$$E(Y \mid X = x) = E(Y \mid X\theta = x\theta), \tag{1}$$

where Y is a scalar-dependent variable, X is a d-dimensional vector of explanatory variables, $x\theta$ is the index, the scalar product of x with θ, a vector of parameters

*Corresponding author.
E-mail addresses: delecroi@ensai.fr (M. Delecroix), haerdle@wiwi.hu-berlin.de (W. Härdle), hristach@ensai.fr (M. Hristache).

0047-259X/03/$ - see front matter
doi:10.1016/S0047-259X(02)00046-5

whose values are unknown. Many widely used parametric models have this form. Examples are linear regression, generalized linear models.

These models assume in (1) a "link" between the index $x\theta$ and the response. In the linear regression, for example, this link is the identity. In the logit model case it is the conditional distribution function of a logistic distribution. In this paper we consider estimation of the parameter θ in (1) without supposing further restrictions on this link. Moreover, we derive the asymptotic normal distribution of this estimator and show that it is efficient in the sense of achieving minimal variance among all estimators for θ.

Several estimators of θ that do not require a fully parametric specification of (1) already exist. Ichimura [12] developed a semiparametric least-squares estimator of θ. This estimator is closely related to projection pursuit regression (see [7]) since it minimizes a least-squares criterion based on nonparametric estimation of the link. Han [9] and Sherman [17] describe a maximum rank correlation estimator. Klein and Spady [13] developed a quasi-maximum likelihood estimator for the case in which Y is a binary response. This estimator achieves the asymptotic efficiency bound of Cosslett [4] if the link is a conditional distribution function. Horowitz and Härdle [10] considered fast noniterative methods for single-index models in the case of discrete covariates. The estimators of [9,10,12,13,17] are $n^{1/2}$-consistent and asymptotically normal under regularity conditions.

Efficient estimation of θ in a single-index model defined by (1) has been considered, for example, in Newey and Stoker [15], based on average derivative estimators (under the assumption of continuous covariates), or Delecroix and Hristache [5], using semiparametric M-estimators. These estimators fail to be efficient in some special single-index regression models, as considered, for example, in [14] or [6], where more information on the conditional law of Y given X is available.

The object of this paper is to construct an asymptotically efficient estimator for general conditional single-index response models, as defined below. Our method will be based on nonparametric estimation of the semiparametric conditional density $f_\theta(y, x\theta)$ of the distribution $\mathscr{L}(Y \mid X\theta)$. We do not assume a specific structure, like for example a binary response as in [13], for this conditional density. Our approach thus covers efficient estimation in linear regression (see [1]) with unknown error distribution as well as nonlinear response models with single-index structure (see [11]). We assume, however, that a continuous conditional density of Y given X exists and has a single-index structure. This allows for using classical nonparametric kernel estimators.

Suppose we are given i.i.d. observations $Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$, with

$$E(Y_i \mid X_i = x) = E(Y_i \mid X_i\theta_0 = x\theta_0), \quad i = 1, \dots, n, \tag{2}$$

where $\theta_0 \in \Theta \subset \mathbb{R}^d$ is the true value of the parameter in the model. Assume that for all $x \in \operatorname{supp} X$ the conditional density $f(y \mid x)$ of Y given $X = x$ with respect to a σ-finite measure exists. This density is supposed to depend upon x through $x\theta_0$. We also assume that the marginal distribution of X does not depend on θ_0.

Thus a positive function f defined on supp $Y \times \mathscr{M}, (\mathscr{M} \subset \mathbb{R})$, is given satisfying:

$$f(y\,|\,x) = f(y, x\theta_0) \quad ((x, y) \in \text{ supp } Z). \tag{3}$$

The main idea of our estimator is to estimate the function f in (3) and then to optimize the (estimated) pseudo-likelihood over the parameter vector θ. The technique is called pseudo-maximum likelihood estimation (PMLE). We use the kernel estimation method here since it is easy to compute in practice and auxiliary asymptotic results are available in the literature. In order to present our estimator we need some more notation.

Let S be a fixed subset of the support of $Z = (X, Y)$ and let $S_X = \{x\colon \exists y \text{ s.t. } (x, y) \in S\}$, $S_Y = \{y\colon \exists x \text{ s.t. } (x, y) \in S\}$, $T_\theta(S) = \{t\colon \exists x \in S_X, \exists \theta \in \Theta$ s.t. $t = x\theta\}$. We assume that for all x in S_X and all θ in Θ, one can define the conditional density $f_\theta(y, x\theta)$ of Y given $X\theta = x\theta$. We will then define the estimator $\hat{f}_\theta$ of f_θ at the point $(y, x\theta)$, for (x, y) in the fixed subset S, by

$$\hat{f}_\theta(y, x\theta) = N_{i,n}(y, x\theta)/D_{i,n}(x\theta) \tag{4}$$

with

$$N_{i,n}(y, t) = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} K_{h_n}(y - Y_j) K_{h_n}(t - X_j\theta),$$

$$D_{i,n}(t) = \frac{1}{(n-1)} \sum_{\substack{j=1 \\ j \neq i}}^{n} K_{h_n}(t - X_j\theta), \tag{5}$$

where h_n is the bandwidth, K is a fixed kernel density, $K_h(\cdot) = K(\cdot/h)/h$.

We define $\hat{\theta}_n$ to be the solution of

$$\hat{L}_n(\hat{\theta}_n) = \max_{\theta \in \Theta} \hat{L}_n(\theta) \tag{6}$$

with

$$\hat{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log \hat{f}_\theta(Y_i, X_i\theta) I_{\{Z_i \in S\}}. \tag{7}$$

Let L_n be the log-likelihood function defined by

$$L_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log f_\theta(Y_i, X_i\theta) I_{\{Z_i \in S\}}. \tag{8}$$

Define also

$$L(\theta) = E\{\log f_\theta(Y_i, X_i\theta) I_{\{Z_i \in S\}}\}. \tag{9}$$

The idea is to maximize the proxy $\hat{L}_n(\theta)$ for $L_n(\theta)$ which itself is a proxy for $L(\theta)$.

2. Consistency of the semiparametric estimator

First, we show that the estimate $\hat{\theta}_n$ defined in (6) converges almost surely to θ_0 as n tends to ∞. We shall prove the following:

$$\sup_{\theta\in\Theta} |\hat{L}_n(\theta) - L_n(\theta)| \xrightarrow[n\to\infty]{\text{a.s.}} 0 \tag{10}$$

and

$$\sup_{\theta\in\Theta} |L_n(\theta) - L(\theta)| \xrightarrow[n\to\infty]{\text{a.s.}} 0. \tag{11}$$

Then, if $L(\theta)$ has a unique maximum at θ_0, the PMLE estimate $\hat{\theta}_n$ converges almost surely towards this maximum. The precise assumptions are as follows:

(A1) $(X_i, Y_i)\in\mathbb{R}^d\times\mathbb{R}$ are i.i.d. random vectors with $E(\|X_i\|^3)<\infty$.
(A2) Θ is a compact subset of $\mathbb{R}^d$.
(A3) The random vectors $(Y_i, X_i\theta)$ have a continuous distribution.
(A4) The compact subset S of the support of $Z_i = (Y_i, X_i)$ is such that:
  (i) for all θ in Θ, the density h_θ of $X_i\theta$ verifies $\inf_{(x,\theta)\in S_X\times\Theta} h_\theta(x\theta)>0$;
  (ii) $\inf_{(z,\theta)\in S\times\Theta} f_\theta(y, x\theta)>0$, where $z = (x, y)$.
(A5) (i) $h_\theta(t), f_\theta(y,t)$, $E(X_i|Y_i = y, X_i\theta = t)$ and $E(X_iX_i^T|Y_i = y, X_i\theta = t)$ are four times differentiable and the fourth-order derivatives satisfy Lipschitz conditions, for $t\in T_\theta(S)$ and $y\in S_Y$, uniformly in $\theta\in\Theta$;
  (ii) $R(x,\theta)\overset{\text{def.}}{=} f_\theta(y, x\theta)$ and $D(x,\theta)\overset{\text{def.}}{=} h_\theta(x\theta)$ are twice continuously differentiable with respect to θ on $S\times\Theta$.
(A6) There exists an unique $\theta_0\in\Theta$ such that relation (3) holds.
(A7) (i) For all θ, $\theta'\in\Theta$ and $x\in S_X$, the distributions P_θ and $P_{\theta'}$ defined by the densities $f_\theta(\cdot, x\theta)$ and $f_{\theta'}(\cdot, x\theta')$ are equivalent.
  (ii) There exists a subset $A\subset S_X$ of positive Lebesgue measure such that X_i is continuous on A.
(A8) The matrix $M = E[-\frac{\partial^2}{\partial\theta\partial\theta^T}\log f_\theta(Y_i, X_i\theta)|_{\theta=\theta_0} I_{\{Z_i\in S\}}]$ is positive-definite.
(C(δ)) K is a real symmetric, compactly supported, twice differentiable fourth-order kernel and $h_n = cn^{-\delta}$, with $c, \delta>0$.

The following preliminary results are shown in the appendix.

Lemma 1. *Under assumptions* (A1)–(A5) *and* (C(δ)) *with* $\delta\in(0,\frac{1}{3})$, *we have*

$$\sup_{z\in S}\sup_{\theta\in\Theta} |\log\hat{f}_\theta(y, x\theta) - \log f_\theta(y, x\theta)| \xrightarrow[n\to\infty]{\text{a.s.}} 0.$$

Lemma 2. *Under assumptions* (A1)–(A5) *and* (C(δ)) *with* $\delta \in (0, \frac{1}{3})$, *we have*

$$\sup_{\theta \in \Theta} |\hat{L}_n(\theta) - L_n(\theta)| + |L_n(\theta) - L(\theta)| \xrightarrow[n \to \infty]{\text{a.s.}} 0.$$

From Lemma 2 it follows that $\sup_{\theta \in \Theta} |\hat{L}_n(\theta) - L(\theta)| \xrightarrow[n \to \infty]{\text{a.s.}} 0$.

Remark. Inspection of the proofs of Lemmas 1 and 2 show that we do not need (A5) in its full strength. Lipschitz continuity is sufficient. For better exposition we use this stronger smoothing throughout.

Lemma 3. *Under assumptions* (A1), (A3) *and* (A5)–(A7) *the function* $L(\theta)$ *has a unique maximum at* θ_0.

The proof relies on the properties of Kullback information and can be found in [3]. Application of [8, p. 431] and Lemmas 1 and 2, yields the following:

Theorem 1. *Under assumptions* (A1)–(A7) *and* (C(δ)) *with* $\delta \in (0, \frac{1}{3})$, *the estimator* $\hat{\theta}_n$ *defined in* (6) *satisfies*:

$$\hat{\theta}_n \xrightarrow[n \to \infty]{\text{a.s.}} \theta_0.$$

3. Asymptotic distribution of the semiparametric estimator

In order to obtain the asymptotic normality of $\hat{\theta}_n$ we show uniform convergence of the first and second derivatives of $\hat{f}_\theta$.

Lemma 4. *Under assumptions* (A1)–(A5) *and* (C(δ)) *with* $\delta \in (\frac{1}{8}, \frac{1}{7})$,

$$n^{1/4} \sup_{(z,\theta) \in S \times \Theta} |\hat{f}_\theta(y, x\theta) - f_\theta(y, x\theta)| \xrightarrow[n \to \infty]{P} 0, \tag{12}$$

$$n^{1/4} \sup_{(z,\theta) \in S \times \Theta} \left| \frac{\partial \hat{f}_\theta(y, x\theta)}{\partial \theta} - \frac{\partial f_\theta(y, x\theta)}{\partial \theta} \right| \xrightarrow[n \to \infty]{P} 0, \tag{13}$$

$$\sup_{(z,\theta) \in S \times \Theta} \left| \frac{\partial^2 \hat{f}_\theta(y, x\theta)}{\partial \theta \partial \theta^T} - \frac{\partial^2 f_\theta(y, x\theta)}{\partial \theta \partial \theta^T} \right| \xrightarrow[n \to \infty]{P} 0. \tag{14}$$

We will show that $-\hat{L}_n(\theta)$ verifies the assumptions of [12, Lemma 5.1]. This is a consequence of Lemma 4 and of the following:

Lemma 5. *Under assumptions* (A1)–(A5)

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{\frac{\partial}{\partial\theta}\log\hat{f}_{\theta}(Y_i, X_i\theta)\bigg|_{\theta=\theta_0}-\frac{\partial}{\partial\theta}\log f_{\theta}(Y_i, X_i\theta)\bigg|_{\theta=\theta_0}\right\}I_{\{Z_i\in S\}}\xrightarrow[n\to\infty]{P} 0.$$

The asymptotic distribution of $\hat{\theta}_n$ is then given by the following:

Theorem 2. *Under assumptions* (A1)–(A8) *and* (C(δ)), $\delta\in(\frac{1}{8},\frac{1}{7})$ *and if* $\theta_0\in\overset{\circ}{\Theta}$, *then*

$$\sqrt{n}(\hat{\theta}_n-\theta_0)\xrightarrow{L}N(0, V), \tag{15}$$

where

$$V=\left\{E\left[\frac{\partial}{\partial\theta}\log f_{\theta}(Y_i, X_i\theta)\bigg|_{\theta=\theta_0}\frac{\partial}{\partial\theta^T}\log f_{\theta}(Y_i, X_i\theta)\bigg|_{\theta=\theta_0}I_{\{Z_i\in S\}}\right]\right\}^{-1}.$$

Proof of Theorem 2. It is sufficient to show that

$$-\frac{1}{n}\sum_{i=1}^{n}\log\hat{f}_{\theta}(Y_i, X_i\theta)I_{\{Z_i\in S\}}$$

verifies conditions (i)–(iv) of [12, Lemma 5.1] of Ichimura.

(i) $\hat{\theta}_n$ converges almost surely to θ_0, by Theorem 1.

(ii) $-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f_{\theta}(Y_i, X_i\theta)|_{\theta=\theta_0}I_{\{Z_i\in S\}}\xrightarrow[n\to\infty]{L}N(0, V)$, since

$$\theta_0=\arg\max_{\theta\in\Theta}E[\log f_{\theta}(Y_i, X_i\theta)I_{\{Z_i\in S\}}]$$
$$\Rightarrow E\left[\frac{\partial}{\partial\theta}\log f_{\theta}(Y_i, X_i\theta)\bigg|_{\theta=\theta_0}I_{\{Z_i\in S\}}\right]=0$$

and

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\partial}{\partial\theta}\log\hat{f}_{\theta}(Y_i, X_i\theta)\bigg|_{\theta=\theta_0}-\frac{\partial}{\partial\theta}\log f_{\theta}(Y_i, X_i\theta)\bigg|_{\theta=\theta_0}\right]\right|I_{\{Z_i\in S\}}$$

converges to 0 in probability by Lemma 5.

(iii) $-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta\partial\theta'}\hat{f}_{\theta}(Y_i, X_i\theta)I_{\{Z_i\in S\}}\xrightarrow[n\to\infty]{P}E[-\frac{\partial^2}{\partial\theta\partial\theta^T}f_{\theta}(Y_i, X_i\theta)I_{\{Z_i\in S\}}]$ uniformly in $\theta\in\Theta$ (Lemma 4 and assumption (A5(ii))).

(iv) $M(\theta_0)=E[-\frac{\partial^2}{\partial\theta\partial\theta^T}f_{\theta}(Y_i, X_i\theta)|_{\theta=\theta_0}I_{\{Z_i\in S\}}]$ is a positive-definite matrix by (A8).

4. Efficiency of the semiparametric estimator

Let $\gamma_0(y,x)$ be the density of $Z_i = (Y_i, X_i)$ (we assume that (Y_i, X_i) is absolutely continuous with respect to a σ-finite measure μ). According to (3) (with f replaced by f_0), for each $z \in S$ we have a decomposition of the form:

$$\gamma_0(y,x) = f_0(y, x\theta_0) g_0(x), \tag{16}$$

where $g_0(x)$ is the marginal density of X_i. Hence, our semiparametric model is defined by the family of distributions

$$\mathscr{P} = \left\{ P\colon \frac{dP}{d\mu} = \gamma(y,x;\theta,f,g),\ \ \theta \in \Theta,\ f \in \mathscr{F},\ \ g \in \mathscr{G} \right\} \tag{17}$$

with the densities γ satisfying:

(i) $\gamma(y,x;\theta,f,g) = f(y,x\theta)g(x)$;
(ii) $\gamma(y,x;\theta_0,f_0,g_0) = \gamma_0(y,x)$.

Following [2], in order to determine the bound of the asymptotic variance of an regular estimator of θ_0, we need to calculate the efficient score. For this purpose, we first need to determine the tangent space $\dot{\mathscr{P}}_2$ corresponding to the nonparametric part

$$\mathscr{P}_2 = \left\{ P\colon \frac{dP}{d\mu} = \gamma(y,x;\theta_0,f,g),\ f \in \mathscr{F},\ \ g \in \mathscr{G} \right\} \tag{18}$$

of the model. This is the closed linear span of the union of tangent spaces corresponding to (one-dimensional) regular parametric submodels $\mathscr{Q} \subset \mathscr{P}_2$. Let

$$\mathscr{Q} = \left\{ P\colon \frac{dP}{d\mu} = \gamma_\eta(y,x;\theta_0) = \gamma(y,x;\theta_0, f(\cdot,\cdot;\eta), g(\cdot;\eta)),\ \ \eta \in \mathscr{H} \subset \mathbb{R} \right\} \tag{19}$$

be such a submodel. Thus $\{f(\cdot,\cdot;\eta)\}_{\eta\in\mathscr{H}} \subset \mathscr{F}$, $\{g(\cdot;\eta)\}_{\eta\in\mathscr{H}} \subset \mathscr{G}$ and there exists an element $\eta_0 \in \mathscr{H}$ such that $\gamma_{\eta_0}(y,x;\theta_0) = \gamma_0(y,x)$. The tangent space $\dot{\mathscr{Q}}$ of $\mathscr{Q}$ (at γ_0) is simply the linear subspace of $L^2(P_0) = L^2(\gamma_0\mu)$ spanned by the score function $S_\eta = \frac{\partial \ln \gamma_\eta(Y_i, X_i;\theta_0)}{\partial \eta}\big|_{\eta=\eta_0}$. We have

$$S_\eta = \left.\frac{\partial \ln f(Y_i, X_i\theta_0;\eta)}{\partial\eta}\right|_{\eta=\eta_0} + \left.\frac{\partial \ln g(X_i,\eta)}{\partial\eta}\right|_{\eta=\eta_0} \tag{20}$$

so that

$$S_\eta \in \mathscr{S} = \{ s_1(Y_i, X_i\theta_0) + s_2(X_i)\colon \\ E_0[s_1(Y_i, X_i\theta_0) \mid X_i\theta_0] = 0,\ \ E_0[s_2(X_i)] = 0 \},$$

where E_0 means that the expectation is taken with respect to the probability measure $P_0 = \gamma_0\mu$. This means that the tangent space $\dot{\mathscr{P}}_2$ is a subspace of $\mathscr{S}$.

Let

$$S_\theta = \frac{\partial \ln \gamma(Y_i, X_i; \theta, f_0, g_0)}{\partial \theta}\bigg|_{\theta=\theta_0} = \frac{\partial \ln f_0(Y_i, X_i\theta)}{\partial \theta}\bigg|_{\theta=\theta_0}$$
$$= \partial_2 \ln f_0(Y_i, X_i\theta_0) X_i. \tag{21}$$

According to [2, Corollary 3.4.1], the information bound on θ_0 is given by $I_0 = E_0(S_\theta^* S_\theta^{*T})$, where S_θ^*, the efficient score, is the residual of the projection of S_θ on $\dot{\mathscr{P}}_2$. Since $\mathscr{S} \subset \dot{\mathscr{P}}_2$, we have $I_0 \geqslant E_0(S_1 S_1^T)$, were $S_1 = S_\theta - \mathrm{proj}(S_\theta | \mathscr{S})$.

On the other hand, if

$$\mathscr{Q} = \left\{ P\colon \frac{dP}{d\mu} = \gamma(y, x; \theta, f(\cdot, \cdot; \theta), g(\cdot; \theta)),\ \ \theta \in \Theta \right\} \tag{22}$$

is a regular parametric submodel of $\mathscr{P}$ containing P_0, then the information bound $I(\theta_0, \mathscr{Q})$ on θ_0 in $\mathscr{Q}$ is such that $I(\theta_0, \mathscr{Q}) \geqslant I_0$. This means that if we can find a parametric submodel $\mathscr{Q}$ such that $I(\theta_0, \mathscr{Q}) = E_0(S_1 S_1^T)$, we have an explicit formula for I_0:

$$I_0 = E_0(S_1 S_1^T)$$
$$= E_0\{[\partial_2 \ln f_0(Y_i, X_i\theta_0)]^2 [X_i - E_0(X_i | X_i\theta_0)][X_i - E_0(X_i|X_i\theta_0)]^T\}, \tag{23}$$

since the projection of $\partial_2 \ln f_0(Y_i, X_i\theta_0) X_i$ is simply $\partial_2 \ln f_0(Y_i, X_i\theta_0) E_0(X_i | X_i\theta_0)$. It is not difficult to see that the submodel:

$$\mathscr{Q} = \left\{ P\colon \frac{dP}{d\mu} = \gamma(y, x; \theta, f_\theta, g_0),\ \ \theta \in \Theta \right\} \tag{24}$$

has the desired property: $I(\theta_0, \mathscr{Q}) = E_0(S_1 S_1^T)$, since

$$\frac{d}{d\theta} f_\theta(Y_i, X_i\theta)\bigg|_{\theta=\theta_0} = \frac{\partial}{\partial t} f_{\theta_0}(Y_i, t)\bigg|_{t=X_i\theta_0} [X_i - E(X_i|X_i\theta_0)]. \tag{25}$$

If we compare I_0 with the asymptotic variance–covariance matrix of our estimator, we can see that, when restricting ourselves to observations Z_i belonging to S, the estimator we proposed is efficient.

5. Simulation study

The asymptotic efficiency of an estimator is not always the most important argument for a practician to use it instead of one which is easy to compute, even if it is not optimal from a theoretical point of view. This is why methods like those proposed by Powell et al. [16] or Horowitz and Härdle [10] are and will be preferred in practice to an estimator which needs optimization procedures, like the one defined by Eqs. (6) and (7). A possible solution to this problem would be to use a one-step estimator, as a compromise between asymptotically and computationally efficiency, whenever we dispose of an estimate easy to compute.

This can be done in the following way: if $\hat{\theta}_n$ is defined by (6) and (7), then $\frac{\partial \hat{L}_n}{\partial \theta}(\hat{\theta}_n) = 0$. If $\tilde{\theta}_n$ is a preliminary $\sqrt{n}$-consistent estimator of θ_0 (we can take, for example, $\tilde{\theta}_n$ as the weighted average derivative estimator of [16]), then we have

$$\frac{\partial \hat{L}_n}{\partial \theta}(\hat{\theta}_n) = \frac{\partial \hat{L}_n}{\partial \theta}(\tilde{\theta}_n) + \frac{\partial^2 \hat{L}_n}{\partial \theta\, \partial \theta^T}(\tilde{\theta}_n)(\hat{\theta}_n - \tilde{\theta}_n) + o_P(||\hat{\theta}_n - \tilde{\theta}_n||).$$

By assumption (A8) and the fact that $\hat{\theta}_n$ and $\tilde{\theta}_n$ are root-n-consistent estimators of θ_0, we obtain:

$$\hat{\theta}_n = \tilde{\theta}_n - \left(\frac{\partial^2 \hat{L}_n}{\partial \theta\, \partial \theta^T}(\tilde{\theta}_n)\right)^{-1} \frac{\partial \tilde{L}_n}{\partial \theta}(\tilde{\theta}_n) + o_P\left(\frac{1}{\sqrt{n}}\right).$$

If we define

$$\bar{\theta}_n = \tilde{\theta}_n - \left(\frac{\partial^2 \hat{L}_n}{\partial \theta\, \partial \theta^T}(\tilde{\theta}_n)\right)^{-1} \frac{\partial \hat{L}_n}{\partial \theta}(\tilde{\theta}_n), \tag{26}$$

we then obtain an asymptotically efficient estimator, since this one-step estimator has the same asymptotic distribution as $\hat{\theta}_n$.

In order to evaluate the performances of the one-step estimator, which is asymptotically equivalent to $\hat{\theta}_n$ but easier to compute, for small sample sizes, we give here the results of a simulation study. We considered the model

$$Y_i = X_i\theta_0 + \varepsilon_i, \quad i = 1, \dots, n,$$

where $Y_i \in \mathbb{R}$, $X_i = (X_i^{(1)}, X_i^{(2)}) \in \mathbb{R}^2$, $\theta_0 = (-1, 1)$, $X_i^{(1)}$ and $X_i^{(2)}$ are independent and of the same law, a mixture of two normal laws,

$$X_i^{(1)}, X_i^{(2)} \sim 0.2\mathcal{N}(0,1) + 0.8\mathcal{N}(0.25, 2),$$

and the errors are normal of mean zero and variance equal to $(X_i\theta_0)^2 = (X_i^{(1)} - X_i^{(2)})^2$:

$$\varepsilon_i \sim N(0, |X_i\theta_0|).$$

As the initial estimator we used the weighted average derivative estimator defined by

$$\tilde{\theta}_n = -\frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{1}{h_n^3} K'\left(\frac{X_i - X_j}{h_n}\right) Y_i,$$

where $K'(\frac{X_i - X_j}{h_n})$ is a notation for the vector

$$\begin{pmatrix} K'(\frac{X_i^{(1)} - X_j^{(1)}}{h_n}) \times K(\frac{X_i^{(2)} - X_j^{(2)}}{h_n}) \\ K(\frac{X_i^{(1)} - X_j^{(1)}}{h_n}) \times K'(\frac{X_i^{(2)} - X_j^{(2)}}{h_n}) \end{pmatrix} \in \mathbb{R}^2,$$

the real-valued kernel function is defined by

$$K(u) = \begin{cases} \frac{1}{4}(7 - 31u^2), & |u| \leqslant \frac{1}{2}, \\ \frac{1}{4}(u^2 - 1), & \frac{1}{2} \leqslant |u| \leqslant 1, \\ 0, & |u| \geqslant 1 \end{cases}$$

and the bandwidth h_n is of the form $h_n = 6n^{-1/5}$.

For the one-step estimator $\bar{\theta}_n$ given by (26), we used the same kernel K in the definition of $\hat{L}_n(\theta)$ and a bandwidth of the form $h_n = 2.5n^{-1/7.5}$.

As only the direction of θ_0 can be identified and not θ_0 itself, we used for the estimators the same constraint as for θ_0, that the last component equals 1. The results for the estimation of $\theta_0^{(1)} = -1$ using the weighted average derivative estimator $\tilde{\theta}_n$ and the one-step estimator $\bar{\theta}_n$ with samples sizes $n \in \{50, 100, 200, 400\}$ are summarized in the following table, containing the empirical mean and the empirical mean squared error for each case:

	$n = 50$	$n = 100$	$n = 200$	$n = 400$
$\tilde{\theta}_n$	-1.0406 (0.1067)	-1.0174 (0.0305)	-1.0187 (0.0152)	-1.0030 (78.45×10^{-4})
$\bar{\theta}_n$	-0.9599 (0.0866)	-0.9649 (0.0269)	-0.9808 (0.0138)	-0.9806 (77.81×10^{-4})

As a general conclusion, we can say that the one-step estimator works better than the initial one. However, the rate of improvement of the squared error decreases with the sample size (18.83%, 11.80%, 9.21% and 0.81%, respectively), but this may be only a consequence of our bandwidth choice, which is for no reason optimal. Moreover, if we change the constant in the bandwidth used to obtain $\bar{\theta}_n$ from 2.5 to 2.0, taking $h_n = 2.0n^{-1/7.5}$, this phenomenon disappears but the general conclusion remains the same, namely that $\bar{\theta}_n$ provides better estimates of θ_0 than $\tilde{\theta}_n$:

	$n = 50$	$n = 100$	$n = 200$	$n = 400$
$\tilde{\theta}_n$	-1.0406 (0.1067)	-1.0174 (0.0305)	-1.0187 (0.0152)	-1.0030 (78.45×10^{-4})
$\bar{\theta}_n$	-0.9812 (0.0961)	-0.9786 (0.0307)	-0.9886 (0.0134)	-0.9857 (75.37×10^{-4})

Acknowledgments

W. Härdle was supported by Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse" Deutsche Forschungsgemeinschaft. This work was substantially completed when M. Delecroix and M. Hristache were visiting at

Sonderforschungsbereich 373. The authors thank V. Spokoiny and J. Horowitz for helpful discussions and comments.

Appendix

Proof of Lemma 1. We first show that

$$\sup_{z \in S} \sup_{\theta \in \Theta} |D_{i,n}(x\theta) - h_\theta(x\theta)| \xrightarrow[n\to\infty]{\text{a.s.}} 0, \tag{A.1}$$

$$\sup_{z \in S} \sup_{\theta \in \Theta} |N_{i,n}(y, x\theta) - f_\theta(y, x\theta)h_\theta(x\theta)| \xrightarrow[n\to\infty]{\text{a.s.}} 0. \tag{A.2}$$

From (A.1) and (A.2) we conclude

$$\begin{aligned}
&\sup_{z \in S} \sup_{\theta \in \Theta} |\log[(N_{i,n}(y, x\theta)/D_{i,n}(x\theta)] - \log f_\theta(y, x\theta)| \\
&\quad \leqslant \sup_{z \in S} \sup_{\theta \in \Theta} |\log N_{i,n}(y, x\theta) - \log[f_\theta(y, x\theta)h_\theta(x\theta)]| \\
&\qquad + \sup_{x \in S_X} \sup_{\theta \in \Theta} |\log D_{i,n}(x\theta) - \log[h_\theta(x\theta)]| \\
&\quad \leqslant \left(\inf_{z \in S} \inf_{\theta \in \Theta} [N_{i,n}(y, x\theta), f_\theta(y, x\theta)h_\theta(x\theta)] \right)^{-1} \\
&\qquad \times \sup_{z \in S} \sup_{\theta \in \Theta} |N_{i,n}(y, x\theta) - f_\theta(y, x\theta)h_\theta(x\theta)| \\
&\qquad + \left(\inf_{x \in S_X} \inf_{\theta \in \Theta} [D_{i,n}(x\theta), h_\theta(x\theta)] \right)^{-1} \sup_{x \in S_X} \sup_{\theta \in \Theta} |D_{i,n}(x\theta) - h_\theta(x\theta)|
\end{aligned}$$

since $\inf_{z \in S} \inf_{\theta \in \Theta} D_{i,n}(x\theta)$ can be asymptotically bounded below almost surely by $\frac{1}{2} \inf_{\substack{x \in S_X \\ \theta \in \Theta}} [h_\theta(x\theta)]$, which is positive by (A4(i)) and similarly $N_{i,n}(y, x\theta)$ is bounded below almost surely by $\frac{1}{2} \inf_{\substack{z \in S \\ \theta \in \Theta}} [f_\theta(y, x\theta)h_\theta(x\theta)] > 0$.

It remains to show (A.1) and (A.2). We show only (A.1). The argument for (A.2) is similar in character.

Since, for $z = (x, y) \in S$, we have

$$D_{i,n}(x\theta) = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} K_{h_n}(x\theta - X_j\theta),$$

adapting results of [12,13] or [18] it suffices to find a function $h_\infty(y, x, \theta)$ such that

$$\sup_{z \in S} \sup_{\theta \in \Theta} |E\{K_h(x\theta - X_j\theta)\} - h_\infty(y, x, \theta)| \xrightarrow[n\to\infty]{} 0.$$

We have, with a big enough constants C', $C>0$,

$$\begin{aligned}
&\sup_{z\in S}\sup_{\theta\in\Theta} |E\{K_h(x\theta - X_j\theta)\} - h_\theta(x\theta)| \\
&\quad\leqslant \sup_{z\in S}\sup_{\theta\in\Theta} \left|\int K_h(x\theta - y)\{h_\theta(y) - h_\theta(x\theta)\}\, dy\right| \\
&\quad\leqslant \sup_{x\in S_X}\sup_{\theta\in\Theta} \int K(u)\{h_\theta(x\theta - uh_n) - h_\theta(x\theta)\}\, du \\
&\quad= \int K(u)\left\{\sup_{x\in S_X}\sup_{\theta\in\Theta} |h_\theta(x\theta - uh_n) - h_\theta(x\theta)|\right\} du \\
&\quad\leqslant C'\left\{h_n \int |u|K(u)\, du\right\} \leqslant Ch_n,
\end{aligned}$$

and then (A.1) is checked.

The proof of Lemma 4 is similar and hence omitted. The only difference is that we use a Taylor expansion in probability.

Proof of Lemma 5. We can write

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \partial_\theta \log \hat{f}_{\theta_0}(Y_i, X_i\theta_0) I_{\{Z_i\in S\}} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\partial_\theta \hat{f}_{\theta_0 i}}{\hat{f}_{\theta_0 i}} I_i.$$

Recalling the definition

$$\hat{f}_\theta(Y_i, X_i\theta) = \frac{N_{i,n}(Y_i, X_i\theta)}{D_{i,n}(X_i\theta)}$$

we have

$$\partial_\theta \hat{f}_{\theta_0 i} = \frac{\partial_\theta N_{i,n}(\theta_0)}{D_{i,n}(\theta_0)} - \hat{f}_{\theta_0 i}\frac{\partial_\theta D_{i,n}(\theta_0)}{D_{i,n}(\theta_0)},$$

which gives

$$\frac{\partial_\theta \hat{f}_{\theta_0 i}}{\hat{f}_{\theta_0 i}} = \frac{\partial_\theta N_{i,n}(\theta_0)}{N_{i,n}(\theta_0)} - \frac{\partial_\theta D_{i,n}(\theta_0)}{D_{i,n}(\theta_0)}$$

and finally

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\partial_\theta \hat{f}_{\theta_0 i}}{\hat{f}_{\theta_0 i}} I_i = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \left(\frac{\partial_\theta N_{i,n}(\theta_0)}{N_{i,n}(\theta_0)} - \frac{\partial_\theta D_{i,n}(\theta_0)}{D_{i,n}(\theta_0)}\right) I_i. \tag{A.3}$$

A decomposition similar to the one used by Ichimura [12] yields

$$\frac{\partial_\theta N_{i,n}(\theta_0)}{N_{i,n}(\theta_0)} = \frac{\partial_\theta g_{\theta_0 i}}{g_{\theta_0 i}} + \frac{1}{g_{\theta_0 i}}\left(\partial_\theta N_{i,n}(\theta_0) - \frac{\partial_\theta g_{\theta_0 i}}{g_{\theta_0 i}} N_{i,n}(\theta_0)\right) + o_p(n^{-1/2}) \tag{A.4}$$

and

$$\frac{\partial_\theta D_{i,n}(\theta_0)}{D_{i,n}(\theta_0)} = \frac{\partial_\theta h_{\theta_0 i}}{h_{\theta_0 i}} + \frac{1}{h_{\theta_0 i}}\left(\partial_\theta D_{i,n}(\theta_0) - \frac{\partial_\theta h_{\theta_0 i}}{h_{\theta_0 i}} D_{i,n}(\theta_0)\right) + o_p(n^{-1/2}). \tag{A.5}$$

Since,

$$\frac{\partial_\theta g_{\theta_0 i}}{g_{\theta_0 i}} = \frac{\partial_\theta (f_{\theta_0 i} h_{\theta_0 i})}{f_{\theta_0 i} h_{\theta_0 i}} = \frac{\partial_\theta f_{\theta_0 i} h_{\theta_0 i}}{f_{\theta_0 i} h_{\theta_0 i}} + \frac{f_{\theta_0 i}\, \partial_\theta h_{\theta_0 i}}{f_{\theta_0 i} h_{\theta_0 i}} = \frac{\partial_\theta f_{\theta_0 i}}{f_{\theta_0 i}} + \frac{\partial_\theta h_{\theta_0 i}}{h_{\theta_0 i}},$$

we obtain

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial_\theta \hat{f}_{\theta_0 i}}{\hat{f}_{\theta_0 i}} I_i = {} & \frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial_\theta f_{\theta_0 i}}{f_{\theta_0 i}} I_i + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial_\theta N_{i,n}(\theta_0)}{g_{\theta_0 i}} I_i}_{\sqrt{n}\, U_{1n}} \\ & - \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial_\theta g_{\theta_0 i}}{g^2_{\theta_0 i}} N_{i,n}(\theta_0) I_i}_{\sqrt{n}\, U_{2n}} - \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial_\theta D_{i,n}(\theta_0)}{h_{\theta_0 i}} I_i}_{\sqrt{n}\, U_{3n}} \\ & + \underbrace{\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial_\theta h_{\theta_0 i}}{h^2_{\theta_0 i}} D_{i,n}(\theta_0) I_i}_{\sqrt{n}\, U_{4n}} + o_p(1).\end{aligned}$$

Writing $U_{1n}, \ldots, U_{4n}$ as second-order U-statistics and applying Lemma 3.1 of [16] yields $\sqrt{n}(U_{kn} - \tilde{U}_{kn}) \xrightarrow[n\to\infty]{P} 0$ $(1 \leqslant k \leqslant 4)$, where

$$\begin{aligned}\tilde{U}_{1n} = {} & \frac{1}{n}\sum_{i=1}^n \left\{ \frac{1}{g_{\theta_0}(Y_i, X_i\theta_0)} [X_i - E(X_i|X_i\theta_0)] \partial_2 g_{\theta_0}(Y_i, X_i\theta_0) I_i \right. \\ & \left. - 2\frac{d}{dt} E(X_i|X_i\theta_0 = t)|_{t=X_i\theta_0} I_i \right\} + E\left[\frac{d}{dt} E(X_i|X_i\theta_0 = t)|_{t=X_i\theta_0} I_i\right],\end{aligned}$$

$$\tilde{U}_{2n} = \frac{1}{n}\sum_{i=1}^n \frac{\partial_\theta g_{\theta_0}(Y_i, X_i\theta_0)}{g_{\theta_0}(Y_i, X_i\theta_0)} I_i,$$

$$\begin{aligned}\tilde{U}_{3n} = {} & \frac{1}{n}\sum_{i=1}^n \left\{ \frac{1}{h_{\theta_0}(X_i\theta_0)} [X_i - E(X_i|X_i\theta_0)] \dot{h}_{\theta_0}(X_i\theta_0) I_i \right. \\ & \left. - 2\frac{d}{dt} E(X_i|X_i\theta_0 = t)|_{t=X_i\theta_0} I_i \right\} + E\left[\frac{d}{dt} E(X_i|X_i\theta_0 = t)|_{t=X_i\theta_0} I_i\right],\end{aligned}$$

$$\tilde{U}_{4n} = \frac{1}{n}\sum_{i=1}^n \frac{\partial_\theta h_{\theta_0}(X_i\theta_0)}{h_{\theta_0}(X_i\theta_0)} I_i.$$

We thus obtain

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial_\theta \hat{f}_{\theta_0 i}}{\hat{f}_{\theta_0 i}}I_i = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial_\theta f_{\theta_0 i}}{f_{\theta_0 i}}I_i + \sqrt{n}(U_{1n} - U_{2n} - U_{3n} + U_{4n}) + o_p(1)$$
$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial_\theta f_{\theta_0 i}}{f_{\theta_0 i}}I_i + \sqrt{n}(\tilde{U}_{1n} - \tilde{U}_{2n} - \tilde{U}_{3n} + \tilde{U}_{4n}) + o_p(1)$$

and finally,

$$\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial_\theta \hat{f}_{\theta_0 i}}{\hat{f}_{\theta_0 i}}I_i - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial_\theta f_{\theta_0 i}}{f_{\theta_0 i}}I_i\right| \xrightarrow[n\to\infty]{P} 0.$$

References

[1] P.J. Bickel, On adaptive estimation, Ann. of Statist. 10 (1982) 647–671.
[2] P.J. Bickel, C.A.J. Klaassen, Y. Ritov, J.A. Wellner, Efficient and Adaptive Estimation for Semiparametric Models, The Johns Hopkins University Press, Baltimore, MD, 1993.
[3] M. Bonneu, M. Delecroix, M. Hristache, Efficient semiparametric estimation in a class of single-index models, Working Paper no. 9447, CREST, Paris, 1994.
[4] S.R. Cosslett, Efficiency bounds for distribution-free estimators of the binary choice and censored regression models, Econometrica 55 (1987) 559–585.
[5] M. Delecroix, M. Hristache, M-estimateurs semi-paramétriques dans les modèles à direction révélatrice unique, Bull. Belg. Math. Soc. 6 (1999) 161–185.
[6] N. Duan, K.-C. Li, Slicing regression: a link-free regression method, Ann. of Statist. 19 (1991) 505–530.
[7] J.H. Friedman, W. Stuetzle, Projection pursuit regression, J. Amer. Statist. Assoc. 76 (1981) 817–823.
[8] C. Gouriéroux, A. Monfort, Statistique et modèles économétriques, Economica, Paris, 1989.
[9] A.K. Han, Non-parametric analysis of a generalized regression model, J. Econom. 35 (1987) 305–316.
[10] J.L. Horowitz, W. Härdle, Direct semiparametric estimation of single-index models with discrete covariates, J. Amer. Statist. Assoc. 91 (1997) 1632–1640.
[11] S. Huet, E. Jolivet, A. Messéan, Some simulations results about confidence intervals and bootstrap methods in nonlinear regression, Statistics 21 (1989) 369–432.
[12] H. Ichimura, Semiparametric least squares (SLS) and weighted SLS estimation of single-index models, J. Econom. 58 (1993) 71–120.
[13] R.L. Klein, R.H. Spady, An efficient semiparametric estimator for binary response models, Econometrica 61 (1993) 387–421.
[14] K.-C. Li, N. Duan, Regression analysis under link violation, Ann. of Statist. 17 (1989) 1009–1052.
[15] W.K. Newey, T.M. Stoker, Efficiency of weighted average derivative estimators and index models, Econometrica 61 (1993) 1199–1223.
[16] J.L. Powell, J.M. Stock, T.M. Stoker, Semiparametric estimation of index coefficients, Econometrica 57 (1989) 1403–1430.
[17] R.P. Sherman, The limiting distribution of the maximum rank correlation estimator, Econometrica 61 (1993) 123–137.
[18] R.P. Sherman, U-processes in the analysis of a generalized semiparametric regression estimator, Econom. Theory 10 (1994) 372–395.